EM for Simple PLDA

Niko Brümmer

November 30, 2010

1 Introduction

This is an EM algorithm for doing a ML estimate of the parameters of a simplified (Gaussian) PLDA model. The PLDA model is simplified by assuming a full (unconstrained) within-class covariance. The between-class covariance can still have low rank.

2 Model

Let observation j of speaker i be $\mathbf{m}_{ij}$ and let it be modelled as:

$$\mathbf{m}_{ij} = \mathbf{V}\mathbf{y}_i + \mathbf{z}_{ij} \tag{1}$$

where

$$\mathbf{y}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{2}$$
$$\mathbf{z}_{ij} \sim \mathcal{N}(\mathbf{0}, \mathbf{D}^{-1}) \tag{3}$$

where the dimension $\mathbf{y}$ may be smaller than that of $\mathbf{m}$ and where $\mathbf{D}$ is the full within-class precision matrix. The *model parameter* that we want to estimate via the EM algorithm is $\boldsymbol{\lambda} = (\mathbf{V}, \mathbf{D})$; and the *hidden variables* are represented by all the $\mathbf{y}_i$. Note that $\mathbf{z}_{ij}$ is not also hidden, because if $\mathbf{m}_{ij}$, $\mathbf{y}_i$ are given, then $\mathbf{z}_{ij}$ is determined.

2.1 Data

We are given N observations of the form $\mathbf{m}_{ij}$, for K speakers, so that $i = 1 \cdots K$. There are n_i observations per speaker, so that $j = 1 \cdots n_i$. We denote the matrix of all the observations for speaker i as $\mathbf{M}_i = \left[\mathbf{m}_{i1} \cdots \mathbf{m}_{in_i}\right]$.

The zero-order statistic for speaker i is n_i and the global zero-order statistic is $N = \sum_{i=1}^{K} n_i$. The first and second-order statistics for speaker i are respectively:

$$\mathbf{f}_i = \sum_{j=1}^{n_i} \mathbf{m}_{ij} \tag{4}$$

$$\mathbf{S}_i = \sum_{j=1}^{n_i} \mathbf{m}_{ij}\mathbf{m}'_{ij} \tag{5}$$

and the global second-order statistic is:

$$\mathbf{S} = \sum_{i=1}^{K} \mathbf{S}_i \,. \tag{6}$$

2.2 Prior

The prior for the hidden variable for a speaker i is:

$$p(\mathbf{y}_i) \propto \exp\left(-\frac{1}{2}\mathbf{y}'_i\mathbf{y}_i\right) \,. \tag{7}$$

2.3 Likelihood

The complete-data likelihood, for speaker i is:

$$p(\mathbf{M}_i|\mathbf{y}_i, \boldsymbol{\lambda}) = \prod_{j=1}^{n_i} \mathcal{N}(\mathbf{m}_{ij}|\mathbf{V}\mathbf{y}_i, \mathbf{D}^{-1}) \tag{8}$$

$$\propto \exp \sum_{j=1}^{n_i} \left(-\frac{1}{2}(\mathbf{m}_{ij} - \mathbf{V}\mathbf{y}_i)'\mathbf{D}(\mathbf{m}_{ij} - \mathbf{V}\mathbf{y}_i) + \frac{1}{2}\log|\mathbf{D}|\right) \tag{9}$$

$$\propto \exp \sum_{j=1}^{n_i} \Big(-\frac{1}{2}\mathbf{m}'_{ij}\mathbf{D}\mathbf{m}_{ij} + \mathbf{m}'_{ij}\mathbf{D}\mathbf{V}\mathbf{y}_i - \frac{1}{2}\mathbf{y}'_i\mathbf{V}'\mathbf{D}\mathbf{V}\mathbf{y}_i + \frac{1}{2}\log|\mathbf{D}|\Big) \tag{10}$$

where factors not dependent on model parameters or hidden variables have been omitted.

2.4 Joint

We express the joint distribution of observed and hidden variables in concise form. Since we use the joint distribution below to derive the posterior for $\mathbf{y}_i$,

we conveniently retain only factors dependent on $\mathbf{y}_i$:

$$p(\mathbf{M}_i, \mathbf{y}_i|\boldsymbol{\lambda}) = p(\mathbf{M}_i|\mathbf{y}_i, \boldsymbol{\lambda})p(\mathbf{y}_i) \tag{11}$$

$$\propto \exp\left(-\frac{1}{2}\mathbf{y}_i'\mathbf{L}_i\mathbf{y}_i + \mathbf{f}_i'\mathbf{D}\mathbf{V}\mathbf{y}_i\right) \tag{12}$$

where

$$\mathbf{L}_i = n_i\mathbf{V}'\mathbf{D}\mathbf{V} + \mathbf{I} \tag{13}$$

2.5 Posterior

The hidden-variable posterior for $\mathbf{y}_i$ is the *key* to the whole EM algorithm. It is found by realizing it is proportional (up to a normalization constant, independent of $\mathbf{y}_i$) to above joint distribution. We can omit all factors in the joint distribution not dependent on $\mathbf{y}_i$. The joint distribution can then be seen to be proportional to a normal distribution, the parameters of which can be recovered by inspection:

$$p(\mathbf{y}_i|\mathbf{M}_i, \boldsymbol{\lambda}) \propto p(\mathbf{y}_i, \mathbf{M}_i|\boldsymbol{\lambda}) \tag{14}$$

$$\propto \exp\left(-\frac{1}{2}\mathbf{y}_i'\mathbf{L}_i\mathbf{y}_i + \mathbf{f}_i'\mathbf{D}\mathbf{V}\mathbf{y}_i\right) \tag{15}$$

$$= \mathcal{N}(\mathbf{y}_i|\hat{\mathbf{y}}_i, \mathbf{L}_i^{-1}) \tag{16}$$

where $\mathbf{L}_i$ is the posterior precision and $\hat{\mathbf{y}}_i$ is the posterior mean, satisfying:

$$\mathbf{L}_i\hat{\mathbf{y}}_i = \mathbf{V}'\mathbf{D}\mathbf{f}_i \tag{17}$$

2.6 Marginal (EM Objective)

The EM objective, which is useful for checking correctness of the implementation and to monitor convergence can be computed conveniently, by making use of the posterior, which we already have:

$$p(\mathbf{M}_i|\boldsymbol{\lambda}) = \left.\frac{p(\mathbf{M}_i|\mathbf{y}_i, \boldsymbol{\lambda})p(\mathbf{y}_i)}{p(\mathbf{y}_i|\mathbf{M}_i, \boldsymbol{\lambda})}\right|_{\mathbf{y}_i=\mathbf{0}} \tag{18}$$

This formula is derived by equating the two ways of factoring the joint distribution and then observing that since the LHS is independent of $\mathbf{y}_i$, we can conveniently set it to zero in the RHS. Expanding the logarithm, we find:

$$\log p(\mathbf{M}_i|\boldsymbol{\lambda}) = -\frac{1}{2}\operatorname{tr}(\mathbf{D}\mathbf{S}_i) + \frac{1}{2}n_i \log|\mathbf{D}| - \frac{1}{2}\log|\mathbf{L}_i| + \frac{1}{2}\hat{\mathbf{y}}_i'\mathbf{V}'\mathbf{D}\mathbf{f}_i + \text{const} \tag{19}$$

The whole objective is the sum over speakers.

3 EM algorithm

In this section we derive formulas for an EM algorithm (with minimum-divergence) for the model described in the previous section. The EM algorithm finds a maximum-likelihood (ML) estimate for the parameter $\boldsymbol{\lambda}$ of the model. We devote subsections to the E-step, the M-step and the (minim-divergence) MD-step.

3.1 EM auxiliary

The EM auxiliary is the expected value (w.r.t. the hidden variable posterior) of the complete data log-likelihood (with irrelevant terms omitted):

$$\tilde{\mathcal{Q}} = \left\langle \sum_i \log p(\mathbf{M}_i|\mathbf{y}_i, \boldsymbol{\lambda}) + \text{const} \right\rangle \tag{20}$$

$$= \left\langle \sum_{ij} \frac{1}{2}\log|\mathbf{D}| - \frac{1}{2}(\mathbf{m}_{ij} - \mathbf{V}\mathbf{y}_i)'\mathbf{D}(\mathbf{m}_{ij} - \mathbf{V}\mathbf{y}_i) \right\rangle \tag{21}$$

$$= \left\langle \sum_{ij} \frac{1}{2}\log|\mathbf{D}| - \frac{1}{2}\mathbf{m}'_{ij}\mathbf{D}\mathbf{m}_{ij} - \frac{1}{2}\mathbf{y}'_i\mathbf{V}'\mathbf{D}\mathbf{V}\mathbf{y}_j + \mathbf{m}'_{ij}\mathbf{D}\mathbf{V}\mathbf{y}_i \right\rangle \tag{22}$$

$$= \frac{N}{2}\log|\mathbf{D}| - \frac{1}{2}\operatorname{tr}(\mathbf{S}\mathbf{D}) - \frac{1}{2}\operatorname{tr}(\mathbf{R}\mathbf{V}'\mathbf{D}\mathbf{V}) + \operatorname{tr}(\mathbf{T}\mathbf{D}\mathbf{V}) \tag{23}$$

where

$$\mathbf{S} = \sum_{ij} \mathbf{m}_{ij}\mathbf{m}'_{ij} \qquad \mathbf{R} = \sum_i n_i \left\langle \mathbf{y}_i\mathbf{y}'_i \right\rangle, \tag{24}$$

$$\mathbf{T} = \sum_i \hat{\mathbf{y}}_i\mathbf{f}'_i, \qquad N = \sum_i n_i. \tag{25}$$

where $\left\langle \mathbf{y}_i\mathbf{y}'_i \right\rangle = \mathbf{L}_i^{-1} + \hat{\mathbf{y}}_i\hat{\mathbf{y}}'_i$.

3.2 M-step

Differentiating w.r.t. $\mathbf{V}$ and setting to zero gives (independently of $\mathbf{D}$):

$$\mathbf{V}' = \mathbf{R}^{-1}\mathbf{T}. \tag{26}$$

Differentiating w.r.t. $\mathbf{D}$, setting to zero and solving gives:

$$\mathbf{D}^{-1} = \frac{1}{N}(\mathbf{S} + \mathbf{V}\mathbf{R}\mathbf{V}' - 2\mathbf{V}\mathbf{T}) \tag{27}$$

$$= \frac{1}{N}(\mathbf{S} - \mathbf{V}\mathbf{T}) \tag{28}$$

$$= \frac{1}{N}(\mathbf{S} - \mathbf{T}'\mathbf{R}^{-1}\mathbf{T}) \tag{29}$$

which is symmetric as required. Note, we used (26) for simplification.

3.3 MD-step

Here we temporarily allow a more general prior for the hidden variables:

$$p(\mathbf{y}_i) = \mathcal{N}(\mathbf{y}_i|\mathbf{0}, \mathcal{Y}), \tag{30}$$

$$\tag{31}$$

and then maximize the following complementary auxiliary w.r.t. to the new prior parameter:

$$\breve{\mathcal{Q}} = \left\langle \sum_i \log \mathcal{N}(\mathbf{y}_i|\mathbf{0}, \mathcal{Y}) \right\rangle \tag{32}$$

This maximization gives:

$$\mathcal{Y} = \frac{1}{K} \sum_{i=1}^{K} \mathbf{L}_i^{-1} + \hat{\mathbf{y}}_i \hat{\mathbf{y}}_i' \tag{33}$$

This non-standard prior can now be transformed back to standard form, by absorbing its effect into $\mathbf{V}$:

$$\mathbf{V} \rightarrow \mathbf{V}\,\mathrm{chol}(\mathcal{Y}) \tag{34}$$

where $\mathrm{chol}(\mathcal{X})\,\mathrm{chol}(\mathcal{X})' = \mathcal{X}$ denotes Cholesky decomposition[1].

[1]Watch out: MATLAB's chol function returns the *transpose* of this definition!